序列标注模型中的字粒度特征提取方案研究*

——以 CCKS2017:Task2 临床病历命名实体识别任务为例

孙安[1,2] 于英香[1] 罗永刚[1,3] 王祺[4]

[1] 上海大学图书情报档案系 上海 200444 [2] 河南科技大学图书馆 洛阳 471023

[3] 上海健康医学院医疗器械学院 上海 201318 [4] 华东理工大学计算机科学与技术系 上海 200237

摘要:［目的/意义］针对中文语言表达特点,提出一种含分词标签的字粒度词语特征提取方法,有效提升了中文临床病历命名实体识别任务的 F_1 值,同时该方法可以为其他中文序列标注模型所借鉴。［方法/过程］选取汉语词语的词性标注、关键词权值、依存句法分析三个特征,构筑字粒度序列标注模型的临床病历训练文本,语料来源 CCKS2017:Task2。在不同特征组合方式下,采用条件随机场算法验证两种字粒度词语特征提取方案 Method1 与 Method2。［结果/结论］在四种不同词语特征组合下,Method2 相对于 Method1 在临床病历命名实体识别任务中性能均有所提升,四折交叉测试中 F_1 值平均提升了 0.23%。实验表明在中文分词技术日趋成熟的环境下,Method2 相对 Method1 能够获得更好的词语特征表示,对中文字粒度序列标注模型的处理性能具有提升作用。

关键词: 命名实体识别 字粒度 特征提取 序列标注模型 条件随机场 临床病历

分类号: TP391.1

DOI: 10.13266/j.issn.0252-3116.2018.11.012

1 引言

临床病历命名实体识别(CNER:Clinical Named Entity Recognition)又称电子病历命名实体识别,它是命名实体识别(NER:Named Entity Recognition)在临床病历文本分析研究中的应用延伸,其任务是利用计算机自动从临床病历文本中识别并抽取出与医学临床相关的命名实体对象,如疾病、症状、检查、治疗等。这些实体对象可供医学临床决策支持等后续医学信息分析研究使用,所以近年来 CNER 的研究发展受到了国内外计算机界、情报界、生物医药界的广泛关注[1]。

真实临床病历语料是 CNER 研究的关键,而临床病历具有私密性,最终由医院归档保存,公开的临床病历语料库非常少。国外研究组织提供的真实临床病历语料可见 I2B2(Informatics for Integrating Biology & the Bedside)-2010[2]、I2B2-2012[3],SemEval(Semantic Evaluation)-2014:Task7[4],SemEval-2015:Task6[5],SemEval-2016:Task12[6],并在这些英文临床病历语料上开展了多项竞赛评测活动。然而由于英文与中文在语言表达方式上具有巨大差异,英文语料处理方法不能完全适用于中文语料。基于此,2017 年中国知识图谱与语义计算大会(CCKS:China Conference on Knowledge Graph and Semantic Computing)围绕"限定领域实体识别与实体链接"这一研究主题开展了中文临床病历命名实体识别竞赛评测活动(CCSK2017:Task2)[7]。本次评测由清华大学知识工程实验室、微软亚洲研究院以及北京极目云健康科技有限公司联合主办,并为该项活动提供了经过脱敏处理的真实临床病历语料集,这也是国内首次以会议组织形式发布的真实中文临床病历语料集。

由于汉语在语言表达上不同于英语,在研究临床病历命名实体识别这类序列标注任务时有"字粒度"和"词粒度"两种方式,大量实验研究表明中文"字粒度"较"词粒度"在序列标注任务中有更好的表现,这

* 本文系国家社会科学基金一般项目"'区域-国家'电子文件管理整合模型构建与实证研究"(项目编号:11BTQ039)研究成果之一。

作者简介:孙安(ORCID:0000-0002-2292-1308),馆员,博士研究生,E-mail:52127688@qq.com;于英香(ORCID:0000-0002-1822-6302),教授,博士生导师;罗永刚(ORCID:0000-0002-8572-335X),讲师,博士研究生;王祺(ORCID:0000-0002-6792-887X),硕士研究生。

收稿日期:2017-10-24 修回日期:2018-03-12 本文起止页码:103-111 本文责任编辑:杜杏叶

是由于“字粒度”为标注模型提供了更多的计算单元，可以得到更多反映实体结构的特征，既能解决部分数据稀疏问题，又能避免分词错误所引入的标注边界错误[8,20,28]。本次 CCSK2017:Task2 所收录的 7 篇评测论文均采用字粒度模型，但并未详细讨论多特征字粒度模型下词语特征的抽取方案，同时词语特征在“字粒度”模型下如何抽取还未见有其他文献详细讨论和实验对比。基于此，本文借鉴联合标签思想，针对中文序列标注任务特点，提出了含分词标签的中文字粒度词语特征提取方法，并选取了词性标注、关键词权值、依存句法三项词语特征，采用多种不同的组合方案进行实验。实验表明，含分词标签的中文字粒度特征提取方法在临床病历命名实体识别中可以获得更好的词语特征表示，四种不同词语特征组合下，Method2 相对 Method1 方法的临床病历命名实体识别的 F_1 值平均提升了 0.23%。实验结果分析中还详细探讨了影响临床病历命名实体识别的其它因素和改进措施。

2 相关工作

临床病历命名实体识别与中文分词、词性标注一样，都可以将其看作是自然语言处理(NLP:Nature Langue Process)研究中的序列标注问题，即给定一个文本序列 $X = <x_1, \cdots\cdots, x_n>$，目标是识别出序列 X 对应的序列 $Y = <y_1, \cdots\cdots, y_n>$。

2.1 序列标注模型概述

序列标注模型不同于一般分类模型，它更强调序列中对象之间的相互联系，即序列中的对象与它前后位置出现的对象之间存在某种关联。这符合人类对自然语言认知的过程，即人们在理解自然语言文本中的某个字或词时，通常是联系上下文来进行的。也有学者将序列标注模型看作是一种特殊的分类模型，即 $Y = <y_1, \cdots\cdots, y_n>$ 是类别，序列标注模型是对文本序列 $X = <x_1, \cdots\cdots, x_n>$ 进行分类，但与传统分类模型不同，当对 x_i 进行预测时，不是孤立的预测 x_i 对应的 y_i，而是联系上下文 $\{\cdots、x_{i-2}、x_{i-1}、x_{i+1}、x_{i+2}、\cdots\}$ 来预测 $\{y_i\}$，这种预测方式也称为“结构化预测”[9]。这种“结构化预测”使得对某一时刻或者某一位置的输入 x_i 进行预测时能够联系更多相互映衬的上下文结构信息，使得序列标注模型在临床病历命名实体识别中较以往的基于字典(Dictionary-based method)[10]和基于规则(Rule-based method)的方法[11]有着更好的性能表现。

目前比较流行的序列标注模型算法有两类：基于人工特征的统计机器学习方法与基于词向量表示技术的神经网络模型算法。基于人工特征的统计机器学习方法，常用的序列标注模型有：隐马尔可夫模型(HMM:Hidden Markov Model)[12]、最大熵马尔可夫模型(MEMM:Maximum Entropy Markov Model)[13]、条件随机场模型(CRF:Conditional Random Fields)[14]等，其中 CRF 模型算法克服了观察值之间的独立假设，采用全局归一化，防止陷入局部最优，解决了标注偏置问题，在实践中取得了较好表现，因此被广泛使用。基于词向量表示技术的神经网络序列标注算法，以双向长短期记忆网络模型(Bi-LSTM:Bidirectional Long Short Term Memory)为代表[15]，它是双向循环神经网络模型(Bi-RNN:Bidirectional Recurrent Neural Network)的改进型。Bi-LSTM 模型可以有效保留或删减长远距离的上下文信息，解决了 Bi-RNN 的梯度消失与梯度爆炸问题，但其求解方式是计算局部最优解，而 CRF 模型求解方式是计算全局最优解，因此可以综合两种模型的优点构筑 Bi-LSTM-CRF 模型[16]。中文临床病历命名实体识别的字粒度 Bi-LSTM-CRF 模型如图 1 所示：

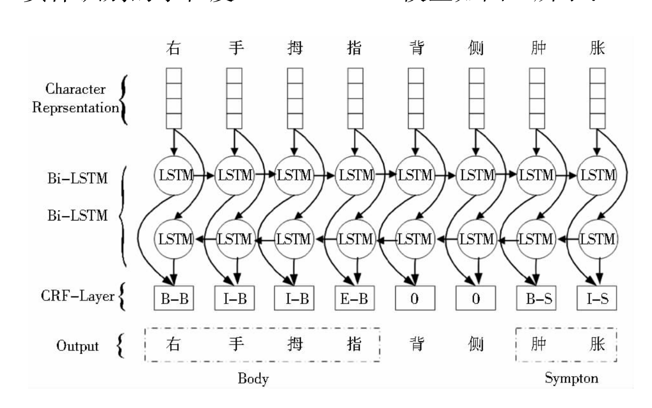


图 1 临床电子病历命名实体识别的 Bi-LSTM-CRF 模型

本次 CCKS2017:Task2 收录的 7 篇评测论文均详细讨论了 Bi-LSTM-CRF(或 Bi-LSTM)模型[17-23]，可见目前受深度学习影响，神经网络模型受到研究者广泛关注。但其中有 3 篇评测论文使用了 CRF 模型与 Bi-LSTM-CRF 模型进行对比[17,22-23]，且 CRF 模型 2 次胜出，可见基于传统人工特征的统计机器学习算法在性能上依然保有竞争力。

2.2 中文序列标注模型中的“字粒度”与“词粒度”

汉语在表达方式上与英语有着巨大差别。英语文本在进行自然语言处理时，其处理的最小单位为英文单词(word)，少有研究基于字母(character)。因为英文单词是最小的语义单元，而字母不具有具体语义，且

英文词与词之间有空格符作为间隔，计算机处理起来也非常方便，所以在构建英文文本的序列标注模型时无字粒度与词粒度之分，基本上都采用“词粒度”模型。

汉语的最小语义单位也是词，由单字、双字或多字构成。但汉语在书写时，词与词之间没有间隔符，所以在中文自然语言处理时，首先对文本进行分词，构建词粒度序列文本成为了很自然的设计思想。文献[24]就是先对文本进行分词，然后利用条件随机场模型构筑中文词粒度的序列标注模型进行关键词自动抽取标引。但由于中文表述的纷繁复杂，仅中文分词就存在粗粒度和细粒度之分[25]，不同粒度的分词结果会对后续任务处理产生影响，且分词结果的误差也会传递给后续任务处理中。同时受中文分词与词性标注采用的是字粒度序列标注模型的影响[26]，人们开始在关键词抽取、命名实体识别等中文自然语言处理任务中，尝试直接采用字粒度方式进行处理。文献[27]详细探讨了基于中文字粒度序列标注模型的关键词提取研究，实验对比了字粒度与词粒度的不同，实验结果表明字粒度处理方法是有效的。文献[8,28]在中文电子病历命名实体识别中也采用了字粒度与词粒度两种序列标注方式进行对比，实验结果显示字粒度要显著优于词粒度。文献[29]在深度神经网络框架下，利用字向量+词向量的拼接方式，获得了中文地名、人名、机构名识别的最好 F_1 值，其方法可以视为中文字粒度序列标注模型思想在神经网络模型中的应用。

2.3 文本特征提取研究

一般认为传统机器学习方法（包括 CRF 算法）的学习效果取决于好的人工特征提取方案，而神经网络模型可以利用表示学习（representation learning）技术和更深的神经网络结构自动学出样本特征的向量表示[30]。但从本次的 7 篇评测论文中可以发现，Bi-LSTM-CRF 这类神经网络模型仅采用单一的字向量（或词向量）并不能获得最佳 F_1 值。利用向量拼接（concatenate）技术，除了引入字向量特征外，还可以增加其他文本特征向量，将这些特征向量与字向量（或词向量）拼接在一起，形成一个高维向量喂给神经网络模型，往往能够取得更好的 F_1 值。7 篇评测论文中用到的向量表示技术，除了分布式表示外，还有独热表示、随机向量表示等方法，见图 2。

可以预见，未来的命名实体识别任务，不论是采用传统机器学习方法还是神经网络方法，关于文本特征的提取和向量表示技术，还将被持续关注与研究。本

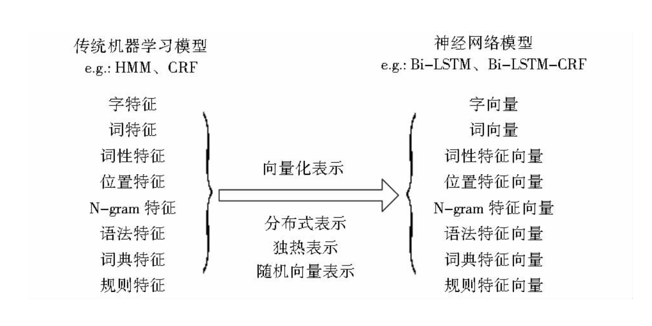


图 2　文本特征设计方法，从传统机器学习模型到神经网络模型

文受联合标签思想启发，在已有字粒度序列标注模型研究基础上，提出一种基于字粒度序列标注模型的词语特征提取方法：将分词的分段标签与该字所在词的词语特征相结合的提取方式，实验证明该方法对中文字粒度序列标注模型是有效的。

3　中文字粒度特征提取方案

3.1　中文单一字符特征序列标注模型

中文单一字符特征序列标注模型中，特征序列为中文汉字，标签序列为标注对象的字粒度表示。标注对象的字粒度表示，一般采用联合标签（cross label）方法实现[31]。联合标签方法是通过将标签集对象的分段标签与命名实体的标签集进行联合生成新的标签集，常见的分段标签集有{B,I,O}、{B,I,O,S}、{B,I,O,E,S}等，具体构造方法和分段标签含义如表 1 所示。通过观察表 1 可以发现，BIOES 联合标签方案较 BIOS 和 BIO 提供了更多的分段信息，识别度更高。文献[32]在 Bi-LSTM 模型下进行实验，验证了 BIOES 编码方案较 BIO 编码方案效果更好。因此，本文论述的字粒度序列标注模型均采用 BIOES 与标签集的联合标签编码模式。

表 1　临床病历文本的 BIO、BIOE、BIOES 三种联合标签字符序列标注模式示意

	患	者	腹	软	，	双	下	肢	无	水	肿	。
BIO-标签	O	O	B-b	B-s	O	B-b	I-b	I-b	O	B-s	I-s	O
BIOS-标签	O	O	S-b	S-s	O	B-b	I-b	I-b	O	B-s	I-s	O
BIOES-标签	O	O	S-b	S-s	O	B-b	I-b	E-b	O	B-s	E-s	O

注：b：身体部位（Body），s：症状体征（Symptom）；B：实体的起始，I：实体的内部，E：实体的结尾，S：单字实体，O：非实体

3.2　字粒度词语特征提取方法（Method1）

序列标注模型单一使用字符特征往往无法取得最佳效果，通常还得挖掘文本的其他语义特征进行联合学习取得最佳效果[33]。语言的最小语义单位是语素，

而汉语语素由词构成(而非字)[34],如:单字词(跑、跳)、双字词(沙发、的士)、多字词(法兰克福、凡士林)。所以汉语文本的语义特征通常是从“词”的角度去分析提取,即提取文本的词语特征。在字粒度序列标注模型中,词语特征提取通常采用“该字所在词的语特征”[28],如表 2 所示。不失一般性,本文从自然语言处理中的三项基本任务:词性标注(Part-of-speech Tagging)、关键词权值提取(Keyword Extraction)、依存句法分析(Dependency parsing)提取词语特征,然后构造临床病历命名实体识别的多特征字粒度序列标注模型。

表 2　“词粒度”与“字粒度”序列标注模型的特征提取方式示意

词粒度	字粒度
词语本身	字本身
词语特征 1	该字所在词的词语特征 1
……	……
词语特征 n	该字所在词的词语特征 n

本文词性标注采用隐马尔可夫算法对中文临床病历文本进行分词并获得词性标注。词性标注结果供关键词权值提取和依存句法分析使用。关键词权值计算:首先对病历文本进行分词,然后采用 TextRank 算法提取关键词,将提取的关键词按照关键词权值从高到低排序,并将其从最高到最低分成 16 个等级,等级作为词语的关键词权值特征,取值范围为{0,1,2,…,15},停用词及其他符号取值为{-1}。TextRank 算法设计思想来源于 Google 的 PageRank 算法,它利用投票原理,让每一个单词给它的邻居投票,票的权重取决于自己的票数,然后采用矩阵迭代收敛的方式获得词的关键词权值的排序[35]。本次临床病历文本词语关键词权值特征提取结果如表 3 所示。依存句法分析最先由法国语言学家 L. Tesniere 于 1959 年提出,它表示的是句子中词语之间的某种依赖关系:一个句子中只有一个成分是独立的,其它成分直接依存于某一成分,任何一个成分都不能依存于两个或两个以上的成分[36]。通过依存句法分析可以获得词的语法成分依存关系。本文采用最大熵模型算法实现病历文本的依存句法分析,图 3 是临床病历文本依存句法分析结果的可视化表示。

表 3　关键词权值提取结果部分示例

权值	关键词提取示例
0	患者、入院、治疗、给予、检查、…
1	感冒、活血化瘀、呕吐物、髋关节、硬膜、…
2	注意、鼻唇沟、撞伤、有力、发红、…
3	胰腺炎、核磁检查、右腕、心影、TCD、…
…	……
13	球棒、稀薄、利多卡因、环状、病情恶化、…
14	未降、稍微、并能、创伤性、发展、…
15	莫于、HGB、钾离子、力散、性反应、…

注:本次在 ccks2017:task 的语料下进行关键词提取,共计提取 7 687 个关键词,除权值 15,其余每个权值下各有 500 个关键词

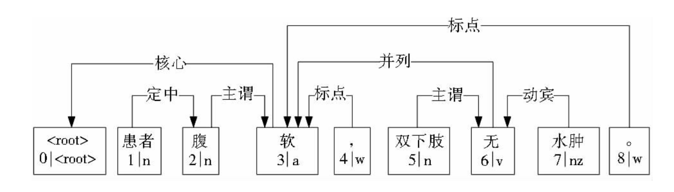


图 3　临床病历文本的依存句法树可视化表示

在获得上述文本词语特征后,字粒度序列标注模型的特征提取方式为“字本身 + 该字所在词的词性特征 + 该字所在词的关键词权值特征 + 该字所在词的依存句法特征”,本节临床病历命名实体识别的训练标注语料如表 4-Method1 所示:

表 4　多特征字粒度序列标注模型临床病例文本训练语料示例

特征列	3.2 节:多特征字粒度模型(Method1)					3.3 节:含分词信息的多特征字粒度模型(Method2)				
	C0	C1	C2	C3	C7	C0	C4	C5	C6	C7
临床病历文本训练语料示例	患	n	0	定中	O	患	B-n	B-0	B-定中	O
	者	n	0	定中	O	者	E-n	E-0	E-定中	O
	腹	ng	-1	主谓	S-Body	腹	S-ng	S--1	S-主谓	S-Body
	软	a	-1	核心	O	软	S-a	S--1	S-核心	O
	,	w	-1	标点	O	,	S-w	S--1	S-标点	O
	双	n	0	主谓	B-Body	双	B-n	B-0	B-主谓	B-Body
	下	n	0	主谓	I-Body	下	I-n	I-0	I-主谓	I-Body
	肢	n	0	主谓	E-Body	肢	E-n	E-0	E-主谓	E-Body
	无	v	-1	并列	O	无	S-v	S--1	S-并列	O
	水	nhd	0	动宾	B-Symptom	水	B-nhd	B-0	B-动宾	B-Symptom
	肿	nhd	0	动宾	E-Symptom	肿	E-nhd	E-0	E-动宾	E-Symptom
		w	-1	标点	O		S-w	S--1	S-标点	O

注:C0:字特征;C1:该字所在词的词性特征;C2:该字所在词的关键词权值特征;C3:该字所在词的依存句法特征;C7:“BIOES”与标注对象的联合标签;C4,C5,C6:“BIES”与 C1,C2,C3 的联合标签

3.3 含分词标签的字粒度特征提取方法(Method2)

上节描述了临床病历文本序列标注模型的字粒度特征提取方案，从表 4 中可以看出，“C1,C2,C3”表示了字所在词的相关词语特征信息，通过分析发现这种字粒度的词语特征提取方案忽略了词语的分词边界信息。以表 4-Method1 示例中的“双下肢”中的“下”字为例，其对应的“n、0、主谓”的三个词语特征，缺乏词的分段边界信息。

本节结合 3.1 节联合标签思想，提出一种字粒度模型下将词语的分词标签(Word-Segment label)与词语特征进行联合的特征提取方案，即字所对应的词语特征为“该字所在词的分词标签—该字所在词的词语特征”。分词标签集为{B、I、E、S}，其中{B、I、E}表示词语的开始、内部、结尾，{S}表示单字词。需要注意的是，此处“BIES”标签的信息来源不同于 3.1 节的“BIOES”标签，前者来源于文本词语分词的分段信息，后者来源于人工标注标签的分段信息。这种含分词分段信息的特征提取方法使得字粒度下的“词特征”信息量更加丰富，不仅表达了字所在词的词语特征，还表达了该特征所在词的位置信息。例如两个并列双字名词序列，在 Method1 方法下，字的词性标注序列是“n，n，n，n”，中间的分词信息将丢失。如果采用 Method2 方法，字的词性标注序列是“B-n，E-n，B-n，E-n”保留了分词信息。这一联合特征提取方法不仅可以在词性标注中运用，还可以运用到其他采用词语粒度计算获得的语义与语法特征，如关键词权值与依存句法特征。本节临床病历命名实体识别的训练标注语料示例如表 4-Method2 所示。

4 实验与分析

4.1 实验任务与数据

本次任务的输入为一组临床病历电子文档，它记录了病人在医院诊断治疗的全过程。任务的输出要求给出文档中与医学相关的命名实体名字的字符串边界，以及每个实体名字对应的类别。本次任务共定义了 5 类命名实体：身体部位(Body)、症状体征(Symptom)、检查检验(Exam)、疾病诊断(Disease)、治疗(Treat)。实验数据为 CCKS2017:Task2 提供的 400 份人工标注数据，人工标签种类分布情况如表 5 所示：

表 5　人工标注的标签种类分布情况

数据集	身体部位	症状体征	检查检验	疾病诊断	治疗	全部
400 份	13 740	10 142	12 689	1 255	1 513	39 359
	34.9%	25.8%	32.3%	3.2%	3.8%	100%

临床病例文本的字粒度特征提取方案采用 3.2 节 Method1 和 3.3 节 Method2，模型的训练采用条件随机场算法(CRF)。实验目的是考察 Method2 方法较 Method1 方法在“临床病历命名实体识别任务”中是否具有性能提升作用。

4.2 实验工具

本次实验平台的编程环境在 Python3.4 与 Jdk1.8 下进行，文本特征提取阶段采用 HanLP 工具。HanLP 是由一系列模型与算法组成的 Java 工具包，通过 Jdk 它可以很方便被 Python 调用。中文分词是 HanLP 的基本功能，也是词性标注、关键词提取、依存句法分析三项功能的基础，文献[18]在与本次实验数据相同环境下对比了 Jieba、NLPIR、Stanford Parser、HanLP 四种分词工具，其中 HanLP 表现最好。多特征字粒度序列标注模型的临床病历文本训练语料生成步骤如图 4 所示：

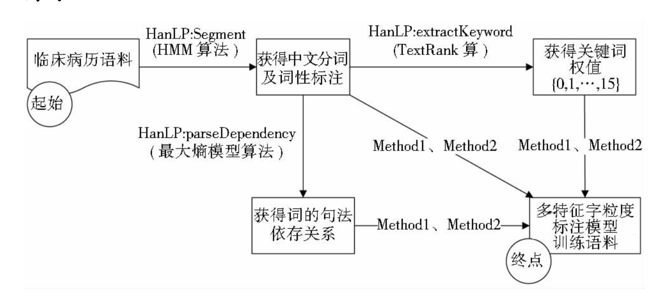


图 4　多特征字粒度序列标注模型训练语料生成流程

条件随机场算法采用 CRF + + 工具实现[37-38]，该工具主要使用：crf_learn 和 crf_test 两个功能。crf_learn 用于训练标注模型，它有多个参数，其中 3 个参数显著影响标注模型的性能：-f，-c，template。在本次实验环境下通过反复多次测试，-f 取 3，-c 取 4.0 可以取得较好性能，该结果与文献[37]研究一致。template 文件定义了条件随机场算法使用的特征模板，如表 6 所示：

表 6　$C_i(i=\{0,1,\cdots,6\})$的特征模板

Ci(i={0,1,⋯,6})的特征模板	
Unigram	U00:%x[-2,i],U01:%x[-1,i],U02:%x[0,i],U03:%x[1,i],U04:%x[2,i]
Bigram	U05:%x[-1,i]/%x[0,i], U06:%x[0,i]/%x[1,i]
Trigram	U07:%x[-2,i]/%x[-1,i]/%x[0,i] U08:%x[-1,i]/%x[0,i]/%x[1,i] U09:%x[0,i]/%x[1,i]/%x[2,i]

注：C_i ∈ 表 4 中的特征列

4.3 实验结果与分析

为了验证 method2 较 method1 方法的有效性，本次

实验在4组不同的词语特征组合方案(Scheme1:字—词性特征、Scheme2:字—词性特征—关键词权值特征、Scheme3:字—词性特征—依存句法特征、Scheme4:字—词性特征—关键词权值特征—依存句法特征)下进行对比测试。

测试过程采用封闭测试和四折交叉测试两种方案。封闭测试(训练集包含测试集),将400份人工标注的电子病历文件作为训练数据,然后将这400份文件分成4组进行标注测试并对结果求平均。四折交叉测试(训练集与测试集不同):将400份人工标注的电子病历文件分成4组,轮流选取3组作为训练数据,剩下1组作为测试数据,然后对测试结果求平均。测试结果评价指标为全体临床病历命名实体识别结果的 F_1 值,以及5类分项命名实体识别结果的 F_1 值。其中:$F_1=2PR/(P+R)$,P为查准率,R为召回率。实验结果如表7所示:

表7 Method1与Method2在不同词语特征组合下命名实体识别的 F_1 值

		Scheme1		Scheme2		Scheme3		Scheme4	
		method1	method2	method1	method2	method1	method2	method1	method2
封闭测试	**F_1(Overall)**	96.65%	**96.67%**	96.57%	**96.67%**	96.95%	**97.01%**	96.96%	**97.02%**
	F_1(body)	94.49%	94.52%	94.35%	94.50%	94.82%	94.91%	94.83%	94.93%
	F_1(Sympton)	98.92%	98.94%	98.86%	98.97%	99.10%	99.11%	99.10%	99.13%
	F_1(Exam)	97.40%	97.44%	97.37%	97.43%	97.60%	97.65%	97.61%	97.67%
	F_1(Disease)	95.25%	95.17%	95.06%	95.28%	96.17%	96.22%	96.26%	96.22%
	F_1(Treat)	96.95%	96.97%	96.96%	96.92%	97.91%	97.95%	97.94%	97.81%
四折交叉测试	**F_1(Overall)**	**89.56%**	**89.82%**	89.48%	**89.61%**	89.04%	**89.40%**	89.13%	**89.31%**
	F_1(body)	84.54%	85.13%	84.54%	84.75%	84.18%	84.79%	84.43%	84.70%
	F_1(Sympton)	95.44%	95.46%	95.36%	95.43%	95.10%	95.16%	95.22%	95.17%
	F_1(Exam)	93.76%	93.54%	93.62%	93.49%	92.99%	93.08%	93.03%	93.10%
	F_1(Disease)	75.05%	73.58%	72.58%	73.22%	72.57%	73.70%	72.61%	72.42%
	F_1(Treat)	74.84%	77.65%	74.20%	76.74%	74.02%	76.07%	72.37	74.83%

对实验结果进行分析:

(1)在Scheme1、Scheme2、Scheme3、Scheme4四种不同的词语特征组合下,不论是四折交叉测试还是封闭测试,总体指标 F_1(Overall)值Method2都要优于Method1,四折交叉验证下 F_1(Overall)平均性能提升0.23%。考察5类实体识别的分项指标 F_1 值,Method2在绝大多数情况下也要优于Method1。通过实例观察:Method2较Method1对单字词实体如"痛—症状体征"、"咽—身体部位"、"肺—身体部位"有更好的识别效果;同时在实体边界识别精度上也有更好表现,如Methiod2下识别为"血常规—检查检验"、"右下腹部髋血管—身体部位",Method1下为"血常规结果"(多字),"右下腹部"(漏字)。实验结果说明通过引入联合标签思想,Method2的特征提取方法较Method1为序列标注模型提供的词语特征信息量更大,不仅提供了词语特征,还提供了该特征所在词的位置信息,对标注效果具有性能增益作用。

(2)由于Method2的特征提取方法依赖于前期词语的分词标签,所以中文分词结果的好坏将影响Method2方法在命名实体识别中性能提升的效果。从理论上分析,如果中文分词结果坏到一定程度,Method2方法对命名实体识别任务的性能增益将会降为负值。但从目前通用中文分词工具的分词结果来看,在临床病历文本中使用Method2方法,性能还是有正增益效果。同时文献[18]提出了一种利用语料集中的人工标签分段信息对中文分词工具的分词结果进行二次矫正的"Re-Segment"分词方法,会有效提高领域文本的分词精度,使用该方法可以进一步提高Method2方法的性能增益。

(3)四折交叉测试中的测试结果要普遍低于封闭测试,尤其Disease与Treat标签,其 F_1 值相对于Body、Sympton、Exam三类标签发生了显著下降,下降幅度超过20%。这说明Disease与Treat标签较Body、Sympton、Exam三类标签存在较严重的过拟合问题。因为在封闭测试时,测试集来自于训练集,产生的误差称为"训练误差"(training error)也叫"经验误差"(empirical error)。训练误差低(F_1 值高)的学习模型不一定是好的学习模型,有些机器学习任务在封闭测试下取得很好的效果,而到实际应用环境中泛化性能下降,效果反而变差,这种现象在机器学习中称为"过拟合"[39]。在交叉测试下,测试集不同于训练集,测试集模拟了新样本,且测试集与训练集的设置比例也不宜"过大"或

“过小”,比例一般在 1/4 ~ 1/2 之间较为合适(四折交叉测试中比例为 1/3),所以四折交叉测试中的 F_1 值较封闭测试更接近模型的实际泛化性能。通过对比封闭测试与四折交叉测试结果可以检验模型是否存在过拟合。

过拟合问题往往是训练集的数据量太少或数据存在噪声等原因造成。通过表 5 发现:本次人工标签的分布比,前三类标签对象 Body、Sympton、Exam 占到总样本数的 93% 。而后两类 Disease 与 Treat 标签仅占总样本数的 3.2% 与 3.8% ,训练数据量明显过小。这使 Disease 与 Treat 标签对象在模型训练时发生的过拟合现象严重。另一方面 Disease 与 Treat 标签对象多为长词结构,识别难度大,而 Body、Sympton、Exam 标签的构词结构相对简单。所以未来临床病历命名实体识别可以通过增加语料方式,扩增 Disease 与 Treat 标签集;同时研究 Disease 与 Treat 长词结构,进一步挖掘词语的语义关系,例如标签集之间的嵌套关系[8],为 Disease 与 Treat 标签词语提供更多的语义特征,提高模型的泛化能力。

(4)数据噪音来源方面分析:四折交叉测试下,随着词特征种类数量的增加,模型的 F_1 值反而下降,模型的泛化能力降低。这说明目前通用领域内关键词特征和依存句法特征提取算法在处理临床病历文本时效果不佳,提取的结果发生错误为序列标注模型的训练引入了噪音。例如临床病历文本中存在大量的简写词与专用词,如“腹部”简写为“腹”,“柔软”简写成“软”,HanLP 的 TextRank 算法将这类单字词作为停用词对待,显然是不合适的,这说明临床医学领域内的关键词提取和依存句法树分析的性能还有待提高。

除词语特征提取存在噪音外,本次测试的人工标注标签也存在一定的噪音。通过人工检查发现,本次训练数据的人工标签存在部分“标签边界二义性”和“标签种类二义性”问题:如“双侧扁桃体”,关于“双侧”有些人工标签包含,而有些则不包含;对于“鼻骨骨折”,有些人工标签标注为“鼻骨”-body、“骨折”-symptom,而有些则标注为“鼻骨骨折”-disease。进一步完善和提高临床病历文本人工标注语料的质量对临床病历文本的信息分析与信息抽取具有重要意义。

5 结语

单特征序列标注模型难以取得最佳效果,多特征联合标注方案将成为主流。随着中文分词技术不断成熟,中文字粒度序列标注模型在词语特征提取上与分词标签进行联合能有效提升标注效果。本文选取了词语的词性标注、关键词权值、依存句法分析三项词语特征进行组合实验研究,实验结果证明了该方法的有效性。同时通过特征的向量化表示技术,该方法得到的词语特征用向量表示后与字向量进行拼接,还可以应用到 Bi-LSTM-CRF 等其他神经网络模型。另一方面,领域语料通常是领域信息分析和数据挖掘的最好研究对象,本次 CCKS2017 提供的真实临床病历语料为临床医学的信息分析与信息抽取提供了鲜活数据,使得本次临床病历命名实体识别的研究结论更具有真实性和实用性。

下一步研究工作:

(1)中文临床病历用语特点不同于一般通用语言,其存在大量的词语简写、语法省略、医学专业术语和受控词汇。过去中文自然语言处理中的基础任务,如分词、词性标注、关键词提取、依存句法分析等在通用语料里取得的较好精度,但却难以满足临床病历文本处理要求。研究开发针对中文临床病历自然语言处理的浅层语义分析工具将对临床病历的信息分析与数据挖掘发挥重要作用。

(2)本次 CCKS2017:Task2 发布的人工标注数据仅提供了临床病历中常见的 5 类命名实体,但相对于国外 I2B2-2010 发布的英文临床病历语料集,还缺少对实体的修饰成分标注和不同类别实体之间的关系标注。例如:“‘无’疼痛”中的‘无’字是一个否定修饰(实体修饰);某种治疗方案施治于特定疾病(实体间关系)。补充完善 CCKS2017:Task2 发布的临床病历标注语料集,引入实体的修饰成分标注和实体关系标注对中文临床科学研究具有重要意义。

参考文献:

[1] 杨锦锋,于秋滨,关毅,等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 40(8):1537－1562.

[2] UZUNER O, SOUTH B R, SHEN S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. Journal of the American medical informatics Association jamia, 2011, 18(5):552－556.

[3] SUN W, RUMSHISKY A, UZUNER O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge[J]. Journal of the American medical informatics Association jamia, 2013, 20(5):806－813.

[4] PRADHAN S, ELHADAD N, CHAPMAN W, et al. SemEval-2014 task 7: analysis of clinical text[C]// Conference: Proceedings of the 8th International workshop on semantic evaluation. SemEval 2014, Dublin, Ireland, 2014:54－62.

[5] BETHARD S, DERCZYNSKI L, SAVOVA G, et al. SemEval-

2015 task 6: clinical tempeval[C]// Conference: proceedings of the 9th International workshop on semantic evaluation. SemEval 2015, Denver, Colorado, 2015:806 –814.

[6] BETHARD S, SAVOVA G, CHEN W T, et al. SemEval-2016 Task 12: clinical tempeval[C]// Conference: Proceedings of the 10th International workshop on semantic evaluation. San Diego, California, 2016:1052 –1062.

[7] CCKS2017 – 全国知识图谱与语义计算大会[EB/OL]. [2017 –08 –26]. http://www.ccks2017.com/.

[8] 王云吉. 基于层叠条件随机场的电子病历命名实体识别[D]. 吉林:吉林大学, 2014.

[9] 汤步洲, 王晓龙, 王轩. 置信度加权在线序列标注算法[J]. 自动化学报, 2011, 37(2):188 –195.

[10] TSURUOKA Y, TSUJII J. Boosting precision and recall of dictionary-based protein name recognition[C]// ACL workshop on natural language processing in biomedicine. Association for computational linguistics, Sapporo, 2003:41 –48.

[11] ALFRED R, LEONG L C, ON C K, et al. Malay named entity recognition based on rule-based approach[J]. International journal of machine learning & computing, 2014, 4(3):300 –306.

[12] LAWRENCE R. RABINER. A tutorial on hidden markov models and selected applications in speech recognition[J]. Readings in speech recognition, 1990, 77(2):267 –296.

[13] BERGER A L, PIETRA V J D, PIETRA S A D. A maximum entropy approach to natural language processing[J]. Computational linguistics, 1996, 22(1):39 –71.

[14] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001). Williamstown: Morgan Kauf-mann, 2001:282 –289.

[15] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural networks the official journal of the international neural network society, 2005, 18(5 –6):602.

[16] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer science, 2015, 20(2):508 –517.

[17] JIANGLU H, XUE S, ZENGJIAN L, et al. HITSZ_CNER: a hybrid system for entity recognition from chinese clinical text[C]// Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). China, 2017:25 –30.

[18] JINHANG W, XIAO H, RONGSHENG Z, et al. Clinical named entity recognition via bi-directional LSTM-CRF model[C]// Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017). China, 2017:31 –36.

[19] OUYANG E, LI Y X, JIN L, et al. Exploring n-gram character presentation in bidirectional RNN-CRF for chinese clinical named entity recognition[C]// Proceedings of the Evaluation Task at the China Conference on knowledge graph and semantic computing (CCKS 2017). China, 2017:37 –42.

[20] XIA Y H, WANG Q. Clinical named entity recognition: ECUST in the CCKS-2017 shared task 2[C]// Proceedings of the evaluation task at the China conference on knowledge graph and semantic computing (CCKS 2017). China, 2017:43 –48.

[21] CHEN Y X, ZHANG G, FANG H Z, et al. Clinical named entity recognition method based on CRF[C]// Proceedings of the evaluation task at the China conference on knowledge graph and semantic computing (CCKS 2017). China, 2017:49 –54.

[22] LI Z Z, ZHANG Q, LIU Y, FENG D W, et al. Recurrent neural networks with specialized word embedding for chinese clinical named entity recognition[C]// Proceedings of the evaluation task at the China conference on knowledge graph and semantic computing (CCKS 2017). Evaluation tasks at CCKS 2017, China, 2017:55 –60.

[23] GENG D W. Clinical name entity recognition using conditional random field with augmented features[C]// Proceedings of the evaluation task at the China conference on knowledge graph and semantic computing (CCKS 2017). China, 2017:61 –68.

[24] 章成志, 苏新宁. 基于条件随机场的自动标引模型研究[J]. 中国图书馆学报, 2008, 34(5):89 –94.

[25] 石崇德, 王惠临. 统计机器翻译中文分词优化技术研究[J]. 现代图书情报技术, 2012, 28(4):29 –34.

[26] 李月伦, 常宝宝. 基于最大间隔马尔可夫网模型的汉语分词方法[J]. 中文信息学报, 2010, 24(1):8 –14.

[27] 王昊, 邓三鸿, 苏新宁. 基于字序列标注的中文关键词抽取研究[J]. 现代图书情报技术, 2011, 27(12):39 –45.

[28] 燕杨, 文敦伟, 王云吉,等. 基于层叠条件随机场的中文病历命名实体识别[J]. 吉林大学学报(工), 2014, 44(6):1843 –1848.

[29] 张海楠, 伍大勇, 刘悦,等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4):28 –35.

[30] 来斯惟. 基于神经网络的词和文档语义向量表示方法研究[D]. 北京:中国科学院自动化研究所, 2016.

[31] 计峰. 自然语言处理中序列标注模型的研究[D]. 上海:复旦大学, 2012.

[32] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[C]// The 55th annual meeting of the association for computational linguistics (ACL). Association for Computational Linguistics . Vancouver, Canada, July 30 - August 4, 2017:1227 –1236.

[33] GU Q, LI Z, HAN J. Joint feature selection and subspace learning [C]// Conference: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, AAAI press, Barcelona, Catalonia, Spain, 2011:1294 –1299.

[34] 柯彼德. 试论汉语语素的分类[J]. 世界汉语教学, 1992(1):1 –12.

[35] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2013, 29(9):30－34.

[36] 唐晓波, 肖璐. 基于依存句法分析的微博主题挖掘模型研究[J]. 情报科学, 2015(9):61－65.

[37] 章成敏, 许鑫, 章成志. 条件随机场标引模型的性能影响因素分析[J]. 现代图书情报技术, 2008(6):34－40.

[38] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016,60(2):122－128.

[39] 周志华. 机器学习: = Machine learning[M]. 北京:清华大学出版社, 2016.

作者贡献说明:

孙安:提出研究思路,制定实验方案,撰写论文初稿;

于英香:设计论文框架,提出修改建议;

罗永刚:为研究选题提供素材和指导;

王祺:提供技术指导。

Research on Feature Extraction Scheme of Chinese-character Granularity in Sequence Labeling Model —— A Case Study About Clinical Named Entity Recognition of CCKS2017:Task2

Sun An[1,2] Yu Yingxiang[1] Luo Yonggang[1,3] Wang Qi[4]

[1] Information and Archival Department, Shanghai University, Shanghai 200444

[2] Library, Henan University of Science and Technology, Luoyang 471023

[3] College of Medical Instrument, Shanghai University of Medicine & Health Sciences, Shanghai 200444

[4] Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237

Abstract: [**Purpose/significance**] According to the characteristics of Chinese language expression, this paper proposes a feature extraction method of words with word segmentation tag of character granularity, which can effectively improve the F_1 value of Chinese clinical named entity recognition, and the method can be used for other Chinese sequence labeling model. [**Method/process**] This paper chose three kinds of features of Chinese-words, including part-of-speech Tagging, keyword weight and dependency parsing, to construct the clinical cases training text in sequence labeling model of the Chinese-character granularity, and the corpus source is CCKS2017:Task2. Then, in different feature combination modes, this paper adopted CRF algorithm to verify Method 1 and Method 2, which are two kinds of words feature extraction methods for character granularity. [**Result/conclusion**] Compared with Method 1, for the four different combinations of word features, Method 2 has been improved in the task of CNER, and the F_1 value has increased by an average of 0.23% in the 4-fold cross-validation test. The experiment shows that in the context of mature Chinese word segmentation technology, Method2 can obtain better word feature representations than Method 1, and it has a lifting effect on the processing performance of Chinese-Character Granularity in Sequence Labeling Model.

Keywords: named entity recognition character granularity feature extraction sequential labeling model conditional random field clinical cases